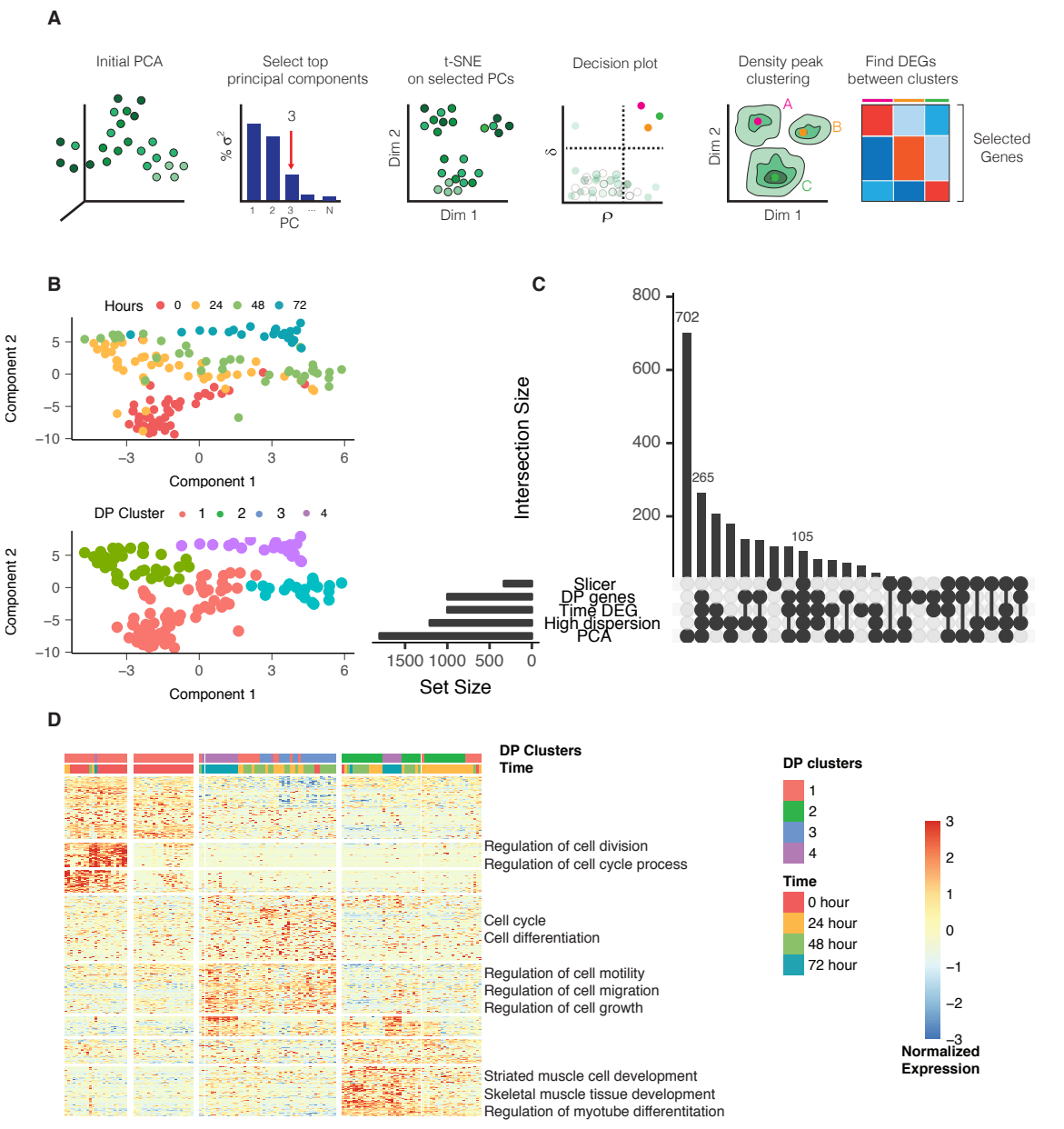
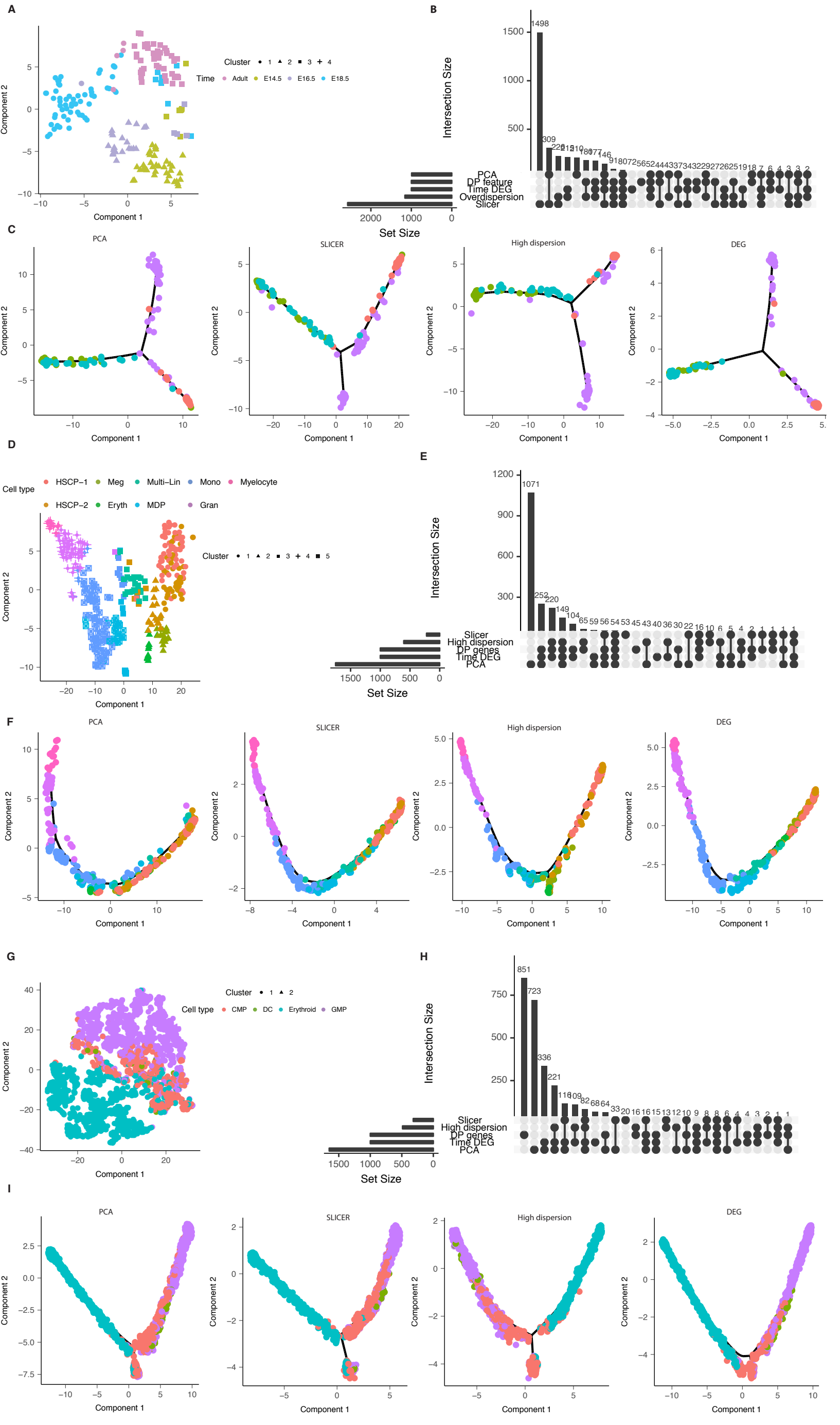
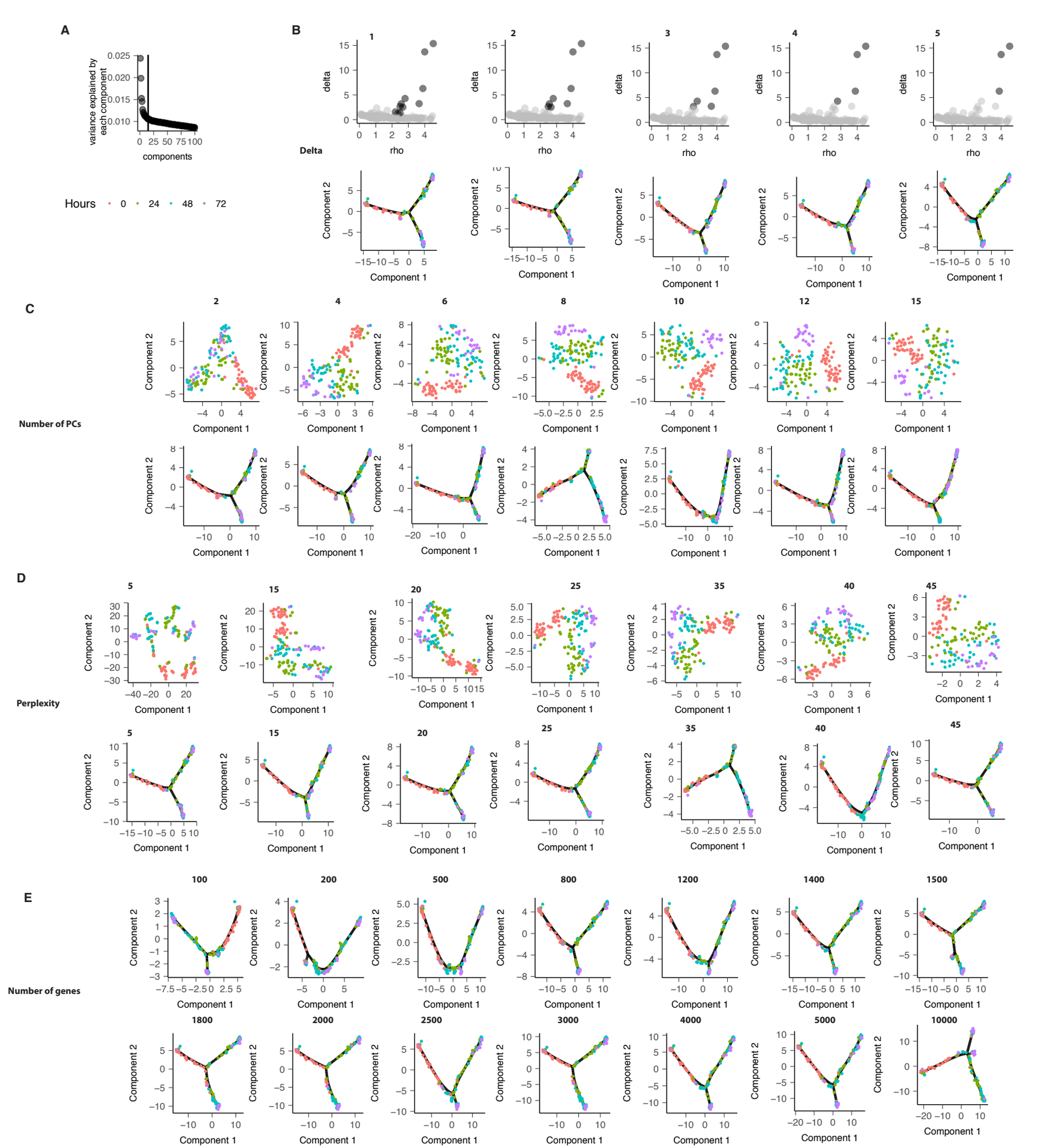


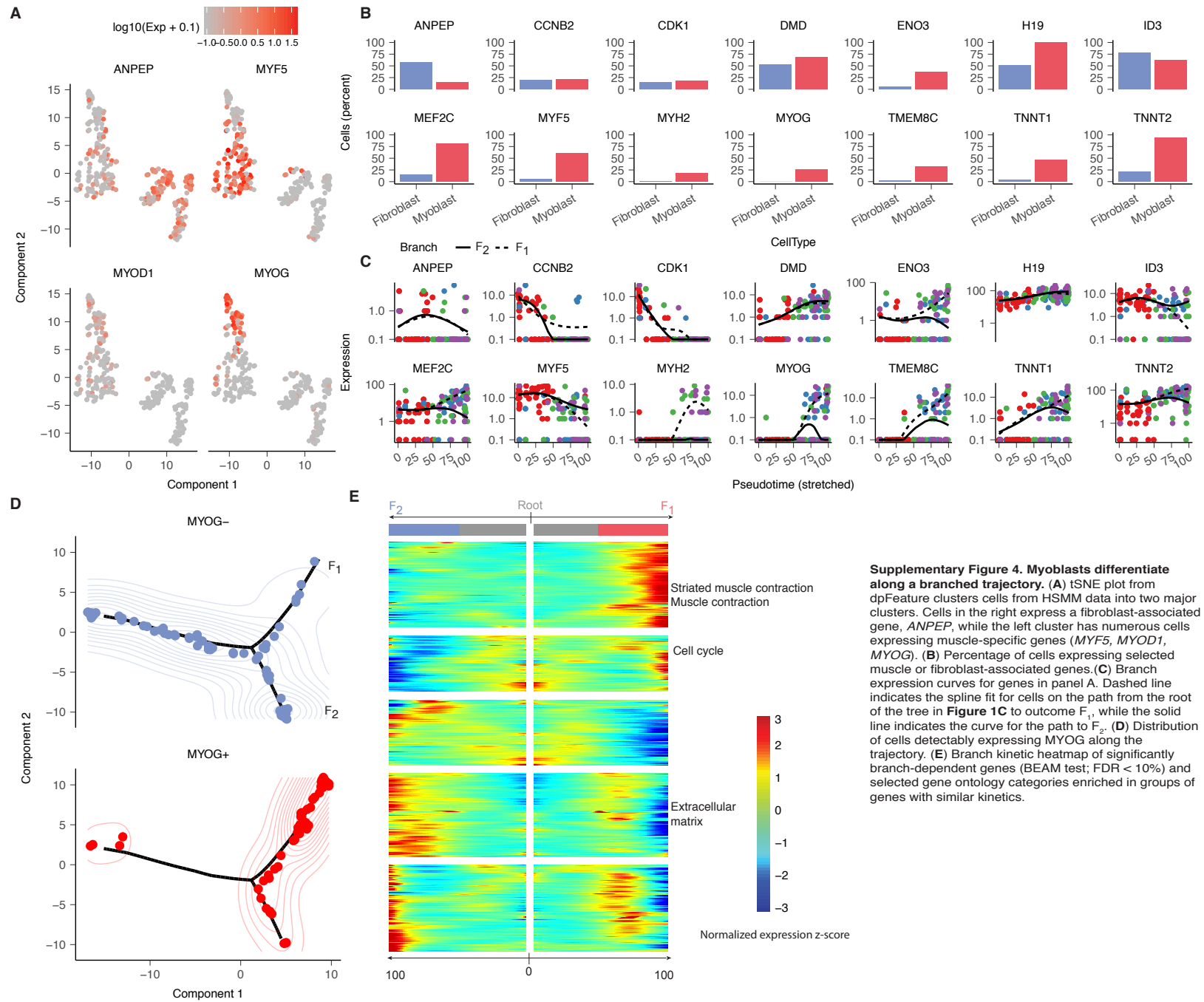
Supplementary Figure 1. dpFeature detects important genes associated with biological processes. (**A**) Flowchart of unsupervised feature selection based on density peak clustering (dpFeature). The density peak algorithm[8] is used to cluster cells in a two-dimensional representation generated by t-SNE. Genes that are significantly differentially expressed between clusters are then selected for downstream trajectory inference in Monocle 2. (**B**) tSNE dimension reduction based on top principal components (PCs) and density peak clustering for the HSMM data. (**C**) Five sets of ordering genes are shown: genes that are highly loaded on the first two principal components ("PCA"), have high dispersion relative to the mean ("High dispersion"), significantly differ between time points ("Time DEG"), were selected via the procedure in SLICER, or identified by dpFeature ("DP genes"). The UpSet plot[32] shows the number of genes returned by each method (bottom left), along with the size of the possible gene set intersections (up right). (**D**) Clustered Heatmap of genes from dpFeature along with selected enriched Gene Ontology terms. Relative transcript counts for each gene (rows) are scaled across cells (columns) and thresholded between the range -3 and 3.



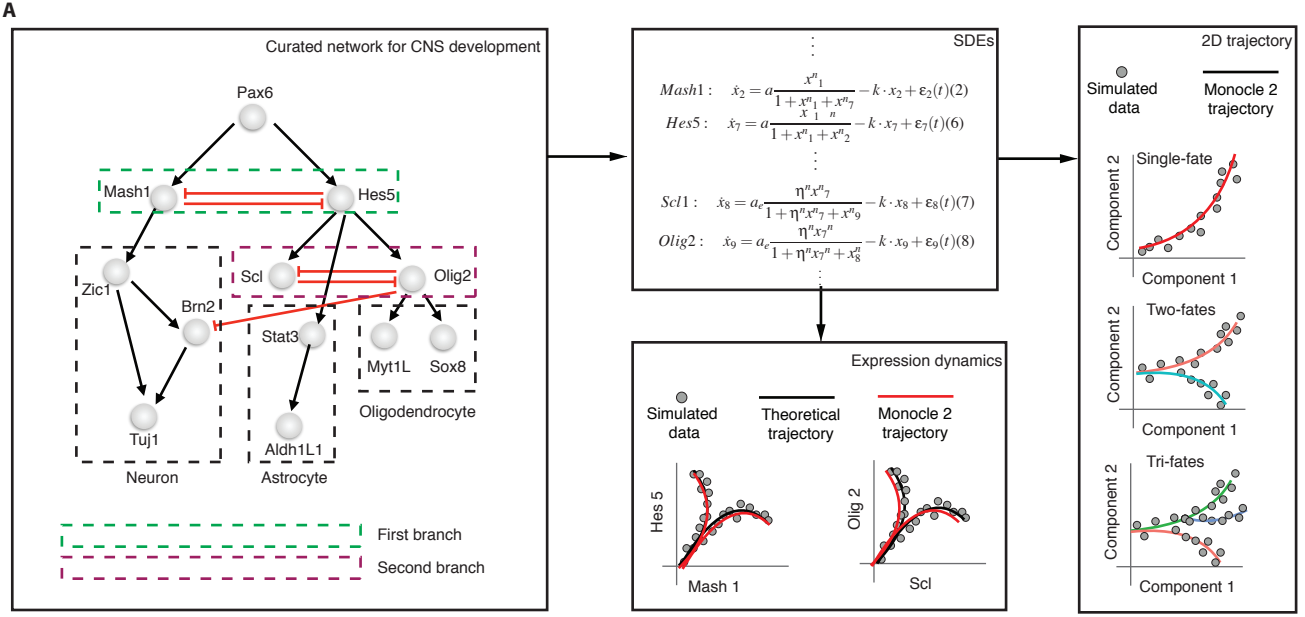
Supplementary Figure 2. DPFeature shapes the reconstruction of developmental trajectories by selecting informative genes. (**A**) tSNE plot from dpFeature clusters cells from lung data[10] into four different clusters. Color corresponds to sample collection time points while shape corresponds to the cluster assignment. (**B**) UpSet plot of the ordering genes selected by various procedures, similar to **Supplementary Figure 1C**. (C) Differentiation trajectories learned with each set of ordering genes. (**D-I**) Similar analysis for the hematopoietic data reported by Olsson et al. (panels **D-F**) and Paul et al. (panels **G-I**).

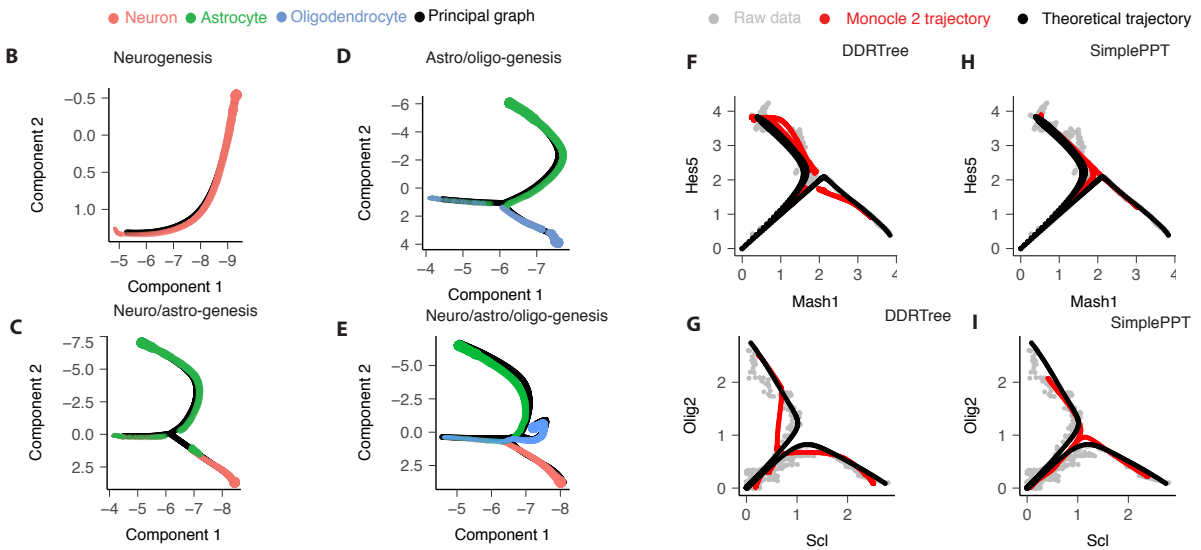


Supplementary Figure 3. Monocle 2 robustly reconstructs the developmental trajectory for HSMM data over a large range of values for parameters used in dpFeature. (**A**) Variance explained by each PCA component for the HSMM dataset. The vertical line corresponds to the 15th PC. The legend for hours is used across all the other panels. (**B-E**) Each panel shows trajectories produced by Monocle 2 while varying one parameter and holding the others fixed at the values specified in the HSMM analysis section (Methods). **B**: Trajectory reconstructed under different values for parameter *delta* used in the density peak clustering step of dpFeature. **C**: Trajectory reconstructed under different value for parameter *number of PC components* used in tSNE dimension reduction step of dpFeature. **D**: Trajectory reconstructed under different value for parameter *perplexity* used for tSNE dimension reduction step. **E**: Trajectory reconstructed under different value for parameter *number of genes* used for trajectory reconstruction.

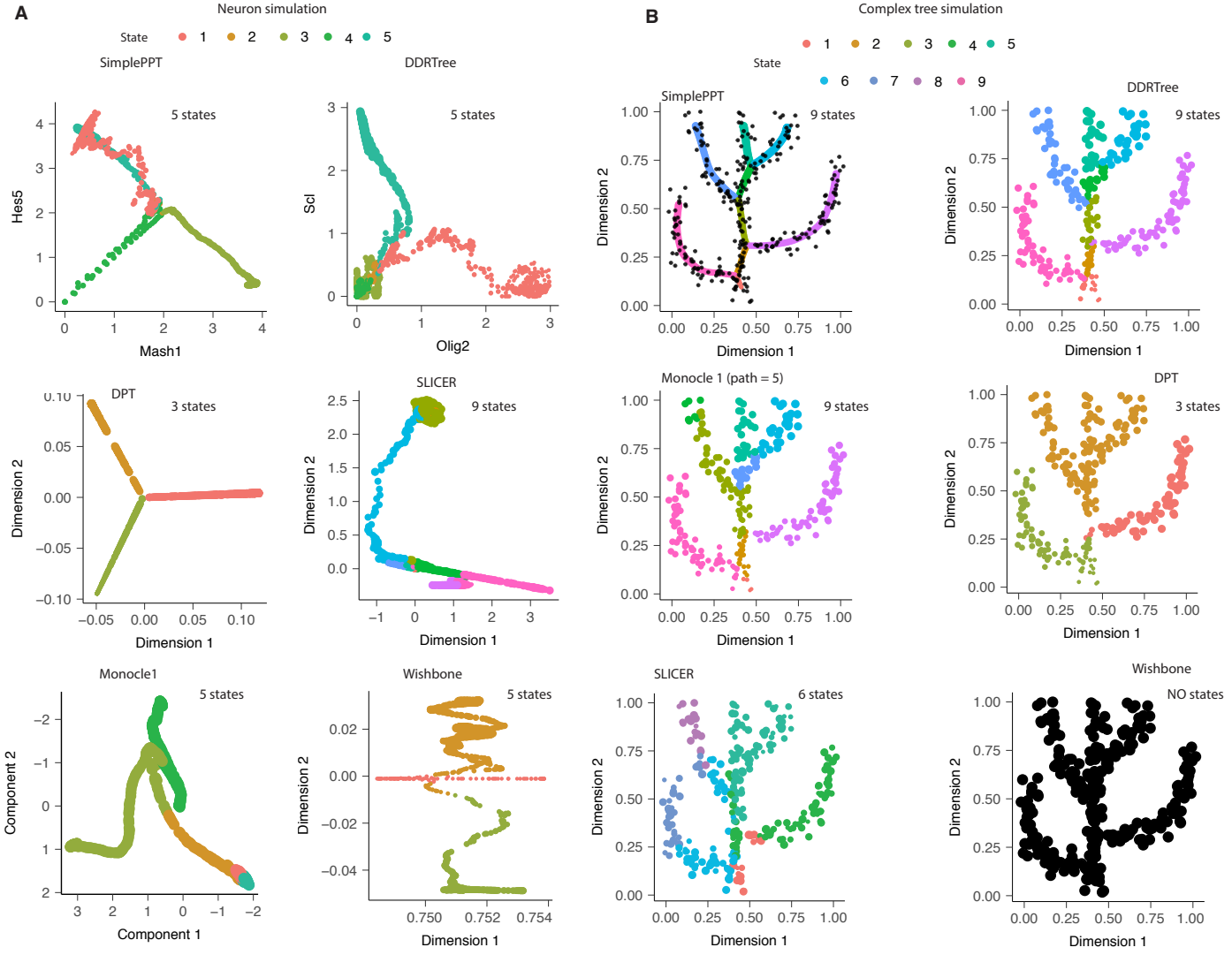


Supplementary Figure 4. Myoblasts differentiate along a branched trajectory. (**A**) tSNE plot from dpFeature clusters cells from HSMM data into two major clusters. Cells in the right express a fibroblast-associated gene, *ANPEP*, while the left cluster has numerous cells expressing muscle-specific genes (*MYF5, MYOD1, MYOG*). (**B**) Percentage of cells expressing selected muscle or fibroblast-associated genes.(**C**) Branch expression curves for genes in panel A. Dashed line indicates the spline fit for cells on the path from the root of the tree in **Figure 1C** to outcome F_1, while the solid line indicates the curve for the path to F_2. (**D**) Distribution of cells detectably expressing MYOG along the trajectory. (**E**) Branch kinetic heatmap of significantly branch-dependent genes (BEAM test; FDR < 10%) and selected gene ontology categories enriched in groups of genes with similar kinetics.

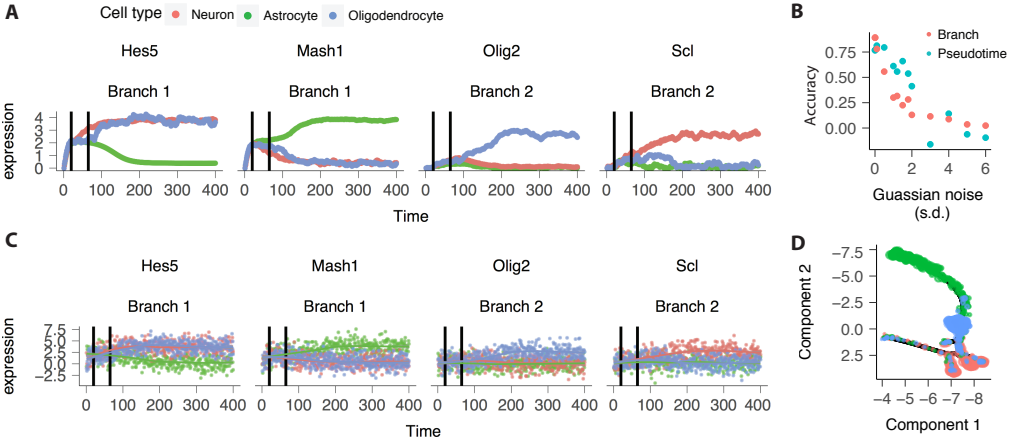




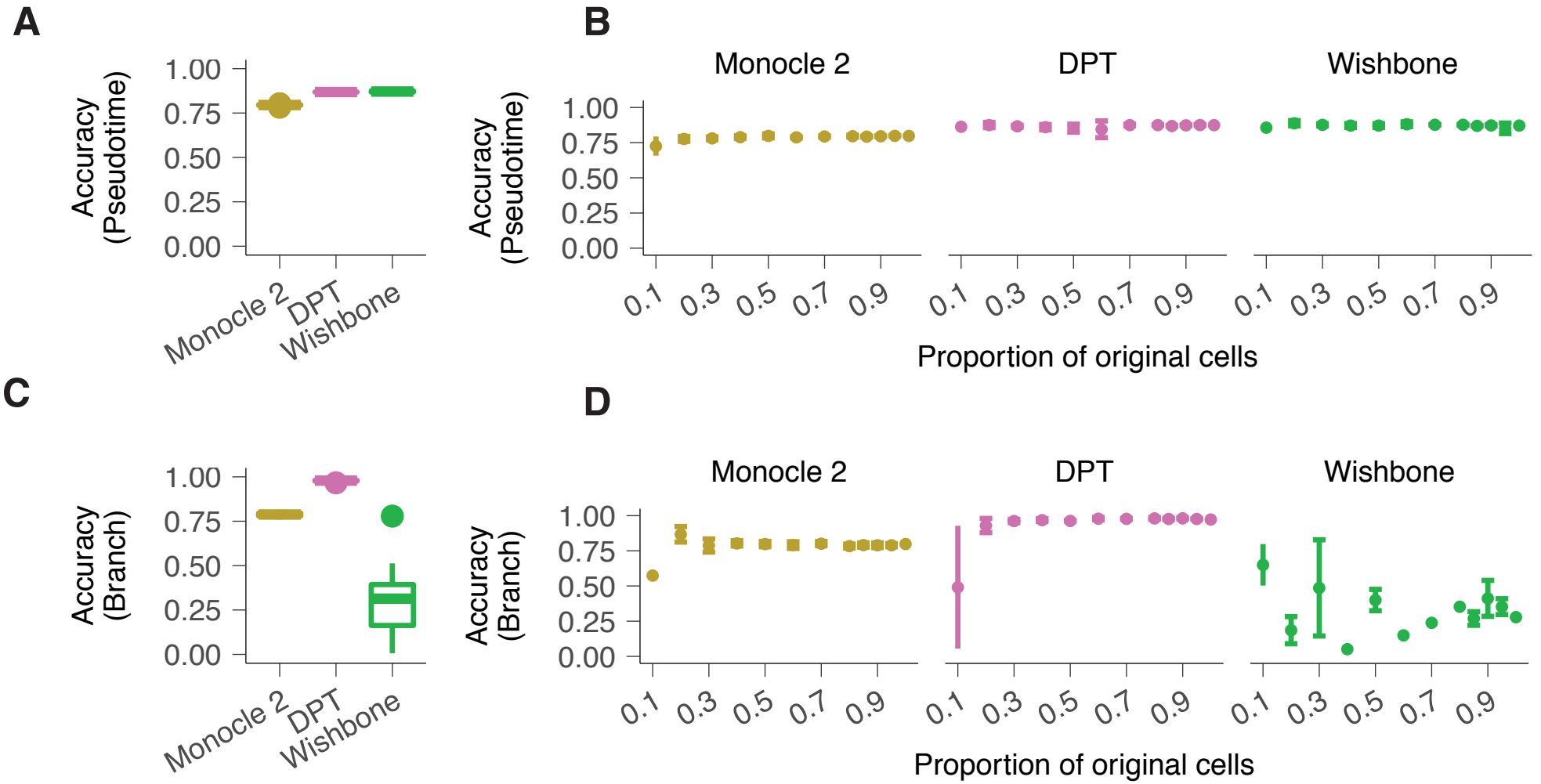
Supplementary Figure 5: Monocle 2 correctly recovers trajectories driven by simulated gene regulatory networks. (**A**) A hypothetical gene regulatory network and system of stochastic differential equations to drive three-way cell fate specification. The transcriptional regulatory network, modified from Qiu et al[33], specifies neural progenitor cells to either neurons, astrocytes or oligodendrocytes through a pair of mutual inhibition interactions[22]. The network summarizes a set of stochastic differential equations that describe gene expression dynamics over time. Initializing this network with small amounts of stochastic noise and following expression kinetics over time simulates the trajectory followed by a single cell, which can be compared to the ideal theoretical trajectory[23]. (**B-E**) Monocle 2 trajectories learned on four ensembles of simulated data points. (**F-I**) Reverse embed (see **Methods**) the lower dimensional principal graph learned by DDRTree back the original gene expression space (**F, G**) or using principal graph from SimplePPT in the same dimension (**H, I**) along with the theoretical trajectory visualizes branching kinetics of individual regulators in the network.



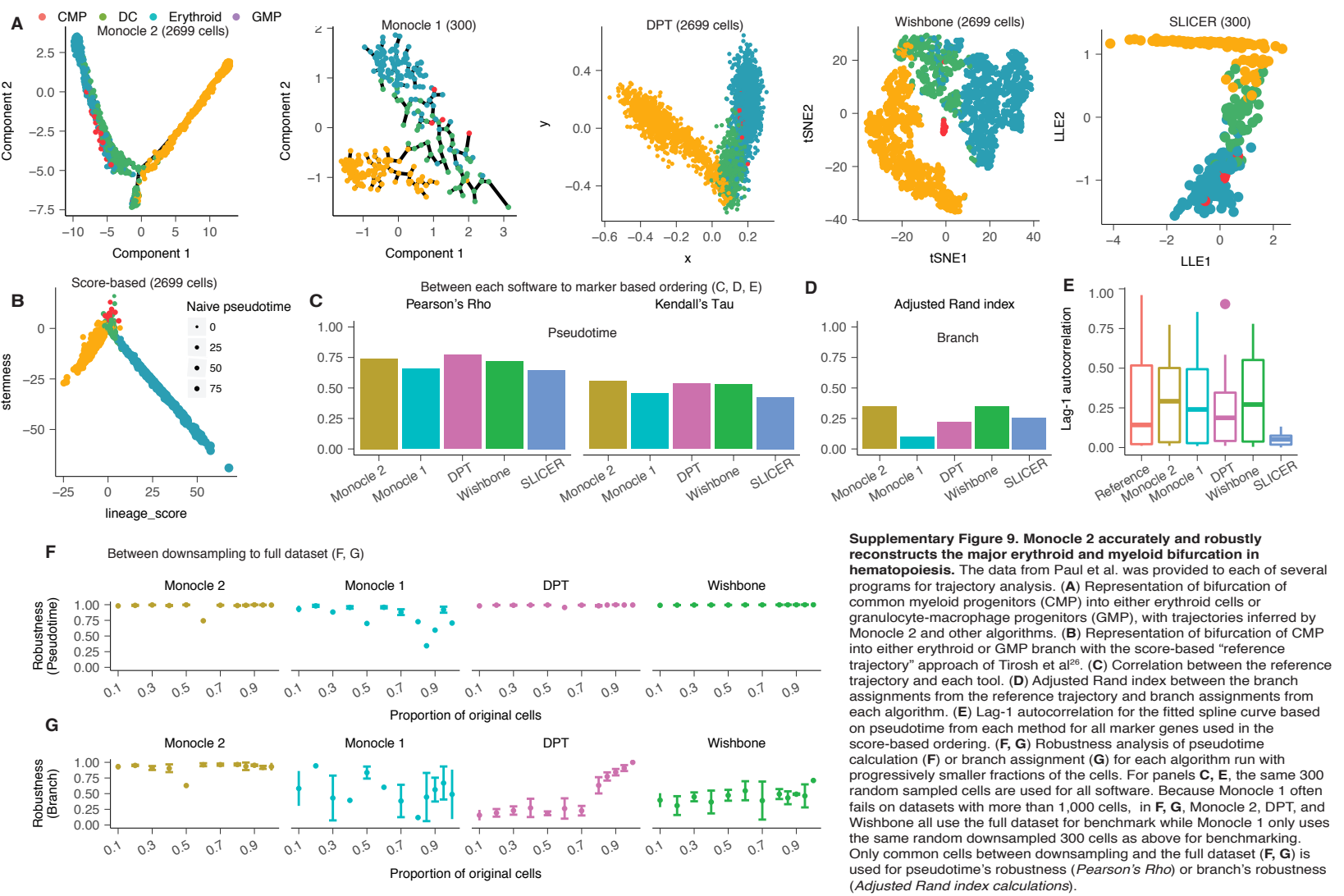
Supplementary Figure 6. Reversed graph embedding accurately learns complex trajectories in simulation datasets. (**A**) Comparison of the trajectory and branch assignment learned by several programs (Monocle 2, either with DDRTree or SimplePPT, DPT, SLICER, Monocle 1, and Wishbone) for the simulated two-branch neuro/astro-oligo-genesis process. (**B**) Comparison of each program for learning a complex tree structure from[20].



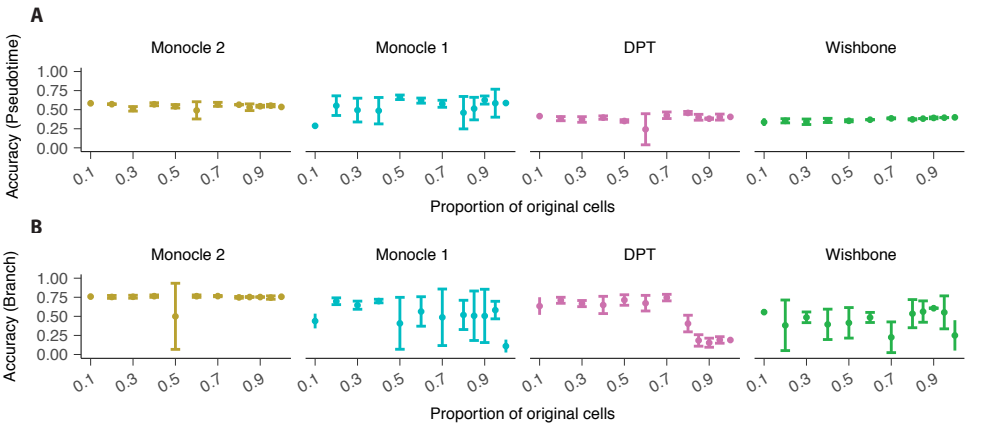
Supplementary Figure 7. Robustness of Monocle 2 over increasing levels of measurement noise in expression data. (**A**) The black vertical lines indicate the branch point pseudotimes in the simulation from **Supplementary Figure 5**. Color represents the cell types. The same colors are used in panels **C** and **D**. (**B**) Accuracy, measured by the pseudotime Pearson correlation and the branch assignment ARI under different levels of Gaussian noise. The x-axis represents the standard deviation of the added Gaussian noise. The real simulation time and branch assignments determined by manual inspection of bifurcation point of master regulator pairs (Mash1-Hes5, Scl-Olig2) in the simulation as shown in panel **A** is used as the reference (also used for **Supplementary Figure 8**). (**C**) Expression dynamics of master regulators under noise level of S.D. of 2. (**D**) Trajectory learned from Monocle 2 under noise level of S.D. of 2.



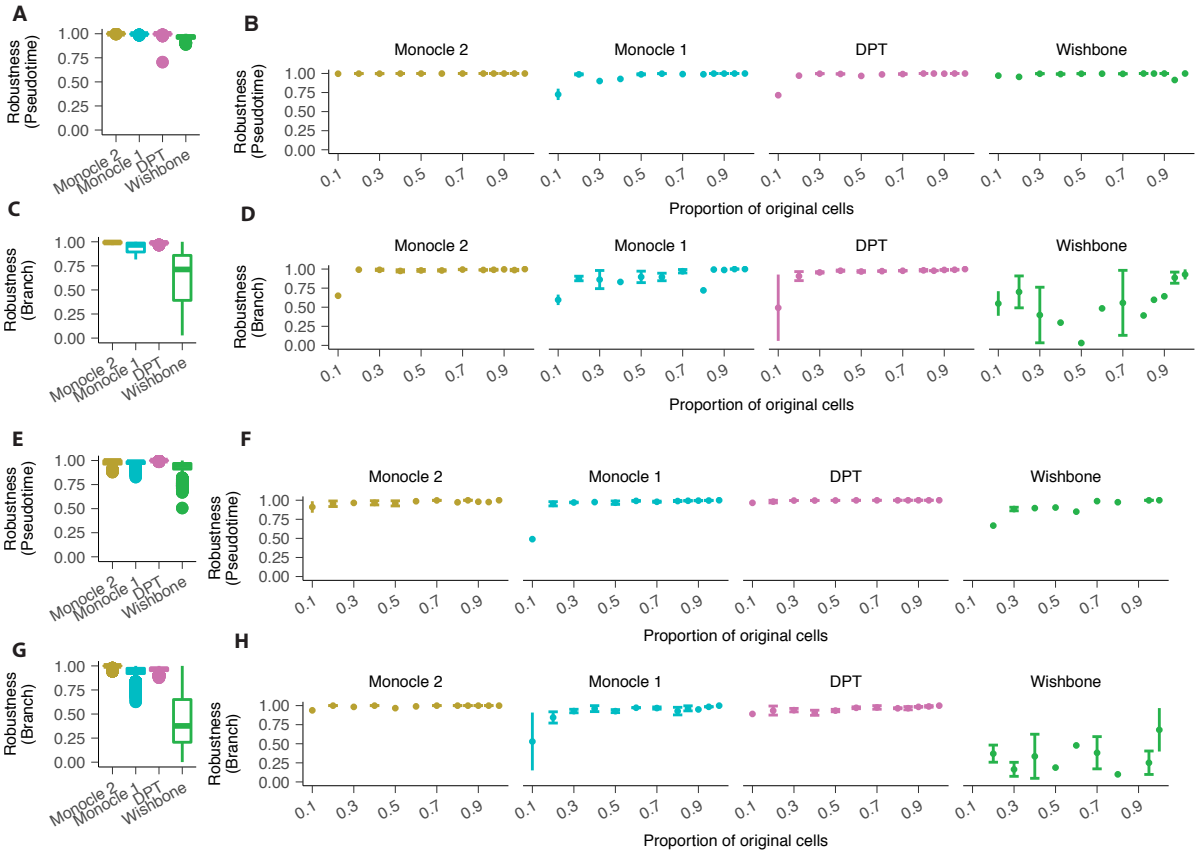
Supplementary Figure 8. Accuracy of trajectory reconstruction algorithms on the simulation dataset. (**A, B**) Accuracy of pseudotime assignment of Monocle 2, DPT, Wishbone over repeated downsampling or progressively smaller fractions of cells from the simulation in Supplementary Figure 6. Pearson correlation between each algorithm's pseudotime assignments and the true simulation times at a depth of 80% of original dataset (**A**) or progressive downsampling (**B**) from 10% to 100% of the full dataset. (**C, D**) Accuracy of branch assignment under repetitive downsampling (**C**) or progressive downsampling (**D**).



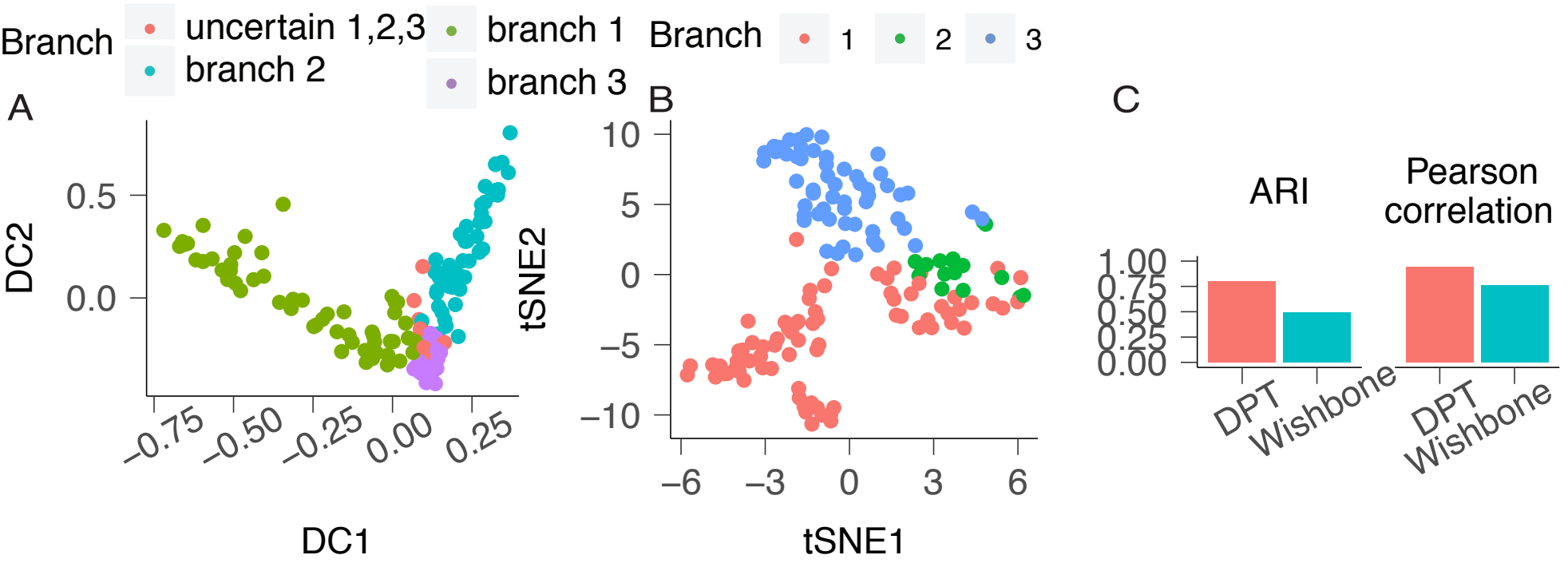
Supplementary Figure 9. Monocle 2 accurately and robustly reconstructs the major erythroid and myeloid bifurcation in hematopoiesis. The data from Paul et al. was provided to each of several programs for trajectory analysis. (**A**) Representation of bifurcation of common myeloid progenitors (CMP) into either erythroid cells or granulocyte-macrophage progenitors (GMP), with trajectories inferred by Monocle 2 and other algorithms. (**B**) Representation of bifurcation of CMP into either erythroid or GMP branch with the score-based "reference trajectory" approach of Tirosh et al[26]. (**C**) Correlation between the reference trajectory and each tool. (**D**) Adjusted Rand index between the branch assignments from the reference trajectory and branch assignments from each algorithm. (**E**) Lag-1 autocorrelation for the fitted spline curve based on pseudotime from each method for all marker genes used in the score-based ordering. (**F, G**) Robustness analysis of pseudotime calculation (**F**) or branch assignment (**G**) for each algorithm run with progressively smaller fractions of the cells. For panels **C, E**, the same 300 random sampled cells are used for all software. Because Monocle 1 often fails on datasets with more than 1,000 cells, in **F, G**, Monocle 2, DPT, and Wishbone all use the full dataset for benchmark while Monocle 1 only uses the same random downsampled 300 cells as above for benchmarking. Only common cells between downsampling and the full dataset (**F, G**) is used for pseudotime's robustness (*Pearson's Rho*) or branch's robustness (*Adjusted Rand index calculations*).



Supplementary Figure 10. Accuracy of trajectory reconstruction algorithms on Paul et. al dataset. (**A**) Same analysis as in **Supplementary Figure 9B** on the Paul dataset with pseudotime from marker based ordering as the reference. (**B**) Same analysis as in **Supplementary Figure 8D** on the Paul dataset with cell type assignment (CMP, GMP or erythroid) suggested by the original study as the reference. For Monocle 1, we used the same random sampled 300 cells from **Supplementary Figure 9** as the universe for benchmark (marker based ordering is also applied on this set of sampled cells).



Supplementary Figure 11. Robustness of trajectory reconstruction algorithms on additional datasets. Robustness of Monocle 2 trajectories for the simulation dataset in **Supplemental Figure 5** (**A-D**) and the lung data from Truetlein et al (**E-H**). Robustness of (**A, E**) pseudotime (Pearson's Rho) or adjusted Rand Index of branch assignments (**C, G**). Values shown are for pairwise comparisons of different downsamples of 80% of the full set of cells. (**B, D**): Robustness of pseudotime assignment (by Pearson correlation) relative to results with the full dataset for Monocle 2, DPT and Wishbone under progressive downsampling from 10% to 100% of the cells. (**D, H**) Same analysis as in (**B, D**) but for robustness of branch assignment (by ARI).



Supplementary Figure 12. Trajectory analysis of myoblast differentiation with alternative approaches. (**A**) DPT's trajectory and branch assignment for the HSMM dataset. The first and second diffusion components are used for visualization. (**B**) Wishbone's trajectory and branch assignment for the HSMM dataset. First and second tSNE components are used for visualization. (**C**) Adjusted Rand index (ARI) for branch consistency and pseudotime's Pearson correlation between branch assignment and pseudotime from DPT, Wishbone to that from Monocle 2.

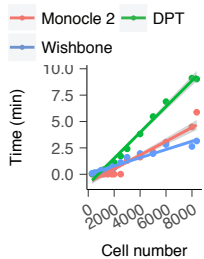
A

	Meaning	Effects of parameters
Dimension	Number of latent space's dimension	Principal tree can be constructed in more than two dimensions to capture more variance of the data
Lambda	Regularization parameter for inverse graph embedding	This parameter controls the length of tree structure. The larger the parameter is, the smaller the length of tree is. In other words, a large lambda will lead to a small tree where points prefer to move to the center of the tree
MaxIter	Maximal number of optimization iterations	The maximum number of iterations is used. The larger it is, the algorithm is more guaranteed to converge. The small MaxIter means early stop and leads to less smooth tree.
Gamma	Regularization parameter for k-means	This parameter controls the fitting of points to its own centers. The larger it is, the better the clusterings appear in the tree structure.
Sigma	Bandwidth parameter	This parameter is used to model the noise of data points. A large sigma is preferred for large noise. The large noise usually causes a large eclipse. A large sigma can enforce points to move to the skeleton of the point cloud that forms the eclipse. In other words, it makes the tree smooth but do not change the main skeleton of the data. A sigma with an overestimated value will shrink the tree structure and make it smoother.
Ncenter	Number of nodes allowed in the regularization graph	The number of clusters used to represent the main structure of the learned tree. It can be same size of the data points if the data size is small. However, if data size is large, it is more efficient to set a moderate number of centers so that the algorithm can run in a reasonable time.

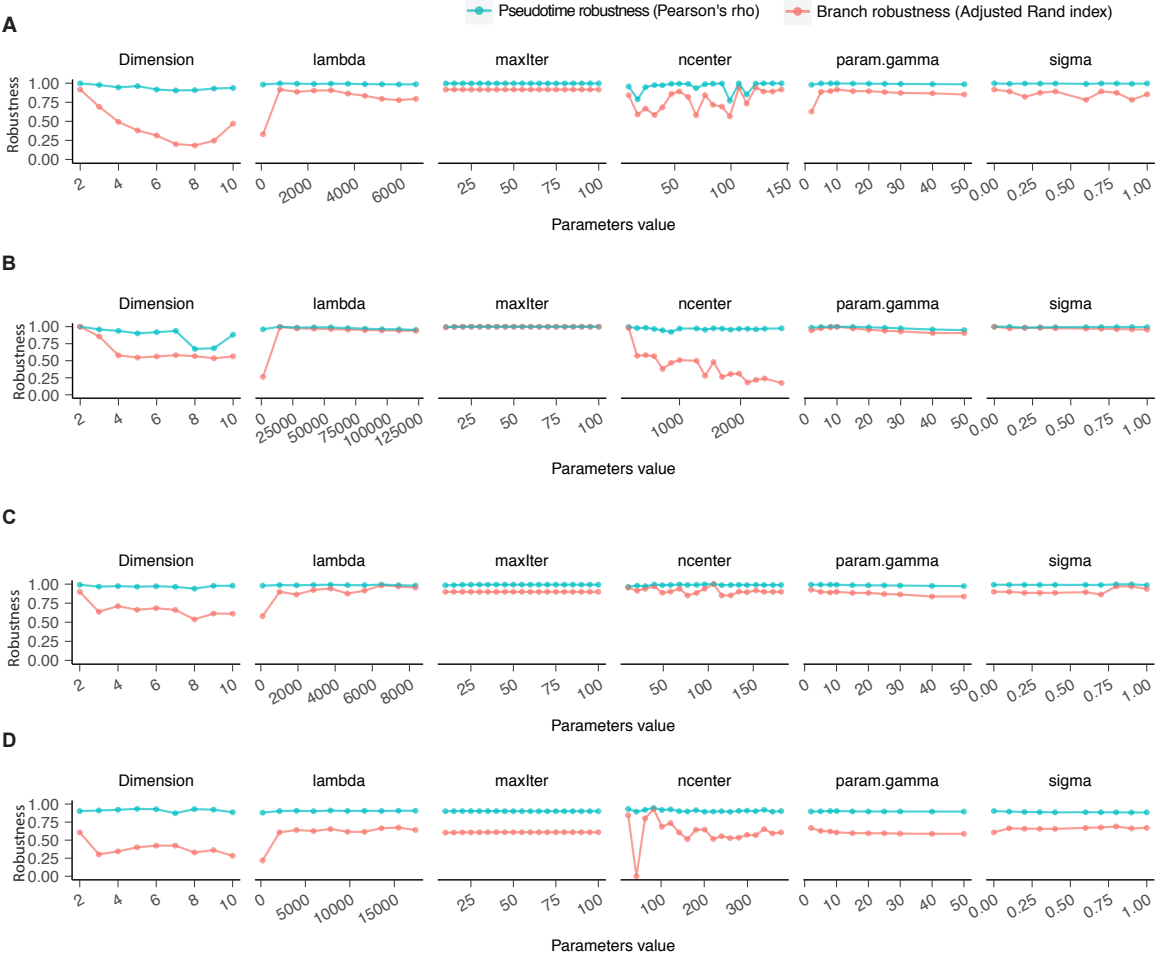
B

Algorithm	Methodology	Tuning parameters	Algorithm complexity (D: gene number; K: centroid number; N: cell number)
DDRTree / DRTree	Learn a set of low-dimensional principal points and a tree over these points	**See above**	$O(K^3+D^3+DK^2+ND^2+NDK)$
SimplePPT	Learning a set of principal points in original space and a tree over these points	**Lambda**: control the balance of the length of tree and the fitness to the input data **Sigma**: smoothness of the tree structure	$O(N^3 + DN^2)$
SGL-tree	Centroids and neighborhood graphs are incorporated into SimplePPT for large-scale problem	**Lambda**: control the balance of the length of tree and the fitness to the input data **Gamma**: fit of the centroids to data **Sigma**: smoothness of the tree structure	$O(K^3+DKN+DK^2)$
L1-graph	Learn a set of principal points in original space and a general sparse graph over these points.	**Lambda**: control the balance of the length of tree and the fitness to the input data **Gamma**: fit of the centroids to data **Sigma**: smoothness of the tree structure	$O(K^3+DKN+DK^2 +L)$ L is the complexity of linear programming

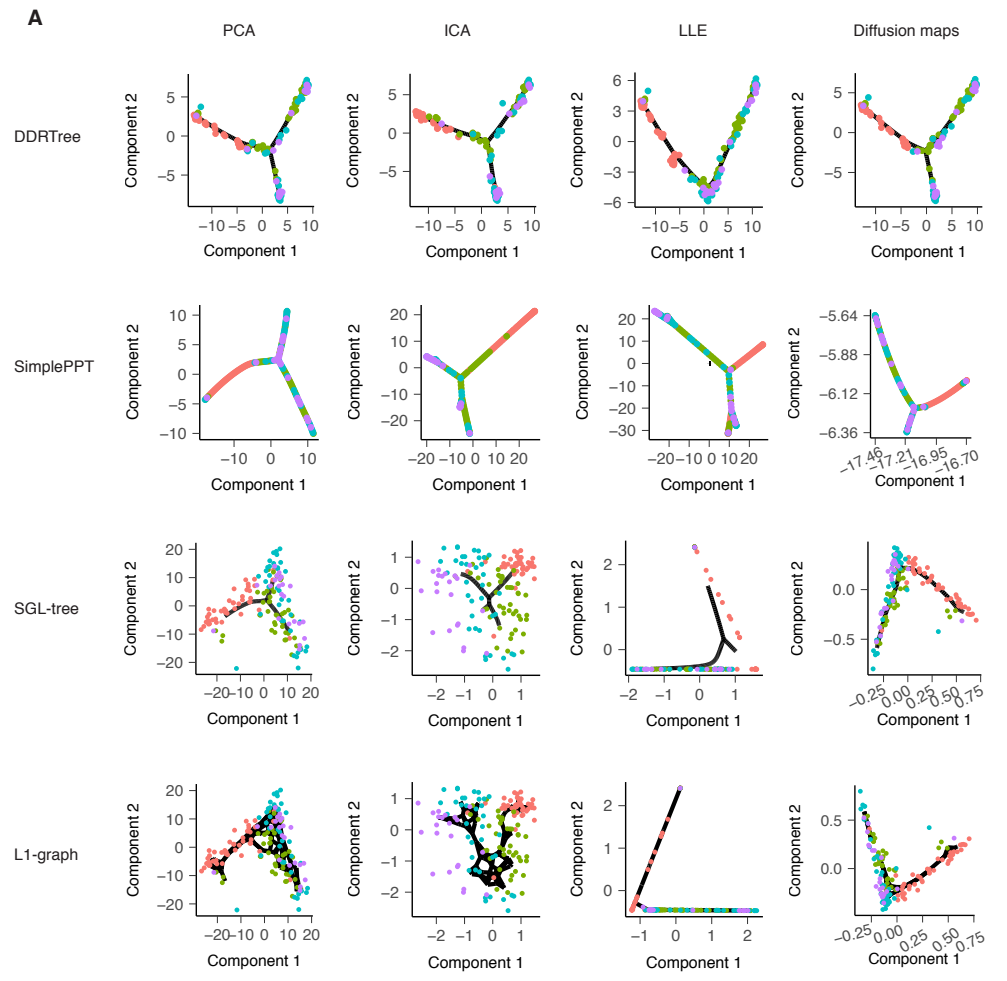
C

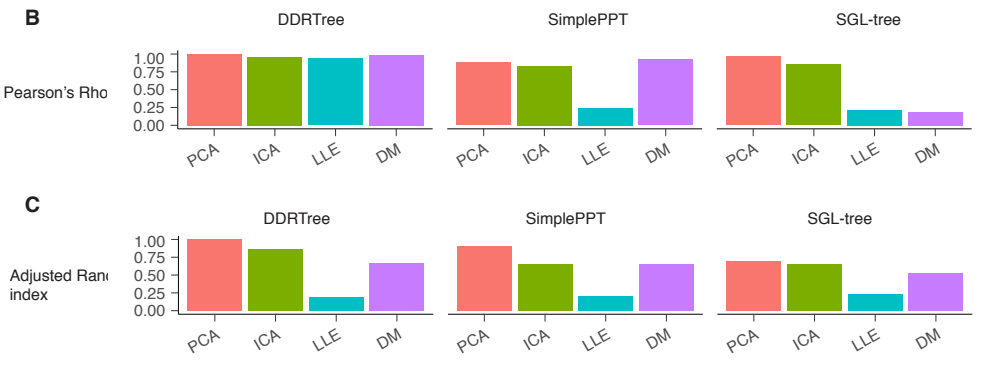


Supplementary Figure 13. Parameters of DDRTree and complexity of current available RGE implementations. (**A**) Parameters used in Monocle 2 (with DDRTree) for trajectory reconstruction. (**B**) Algorithmic complexity for DDRTree, SimplePPT, SGL-tree and L1Graph as a function of cells, ordering genes, and dimension of the embedding space. (**C**) Running time for Monocle 2, DPT and Wishbone on the full dataset (8365 cells) from Paul et al. over varying fractions of the cells.

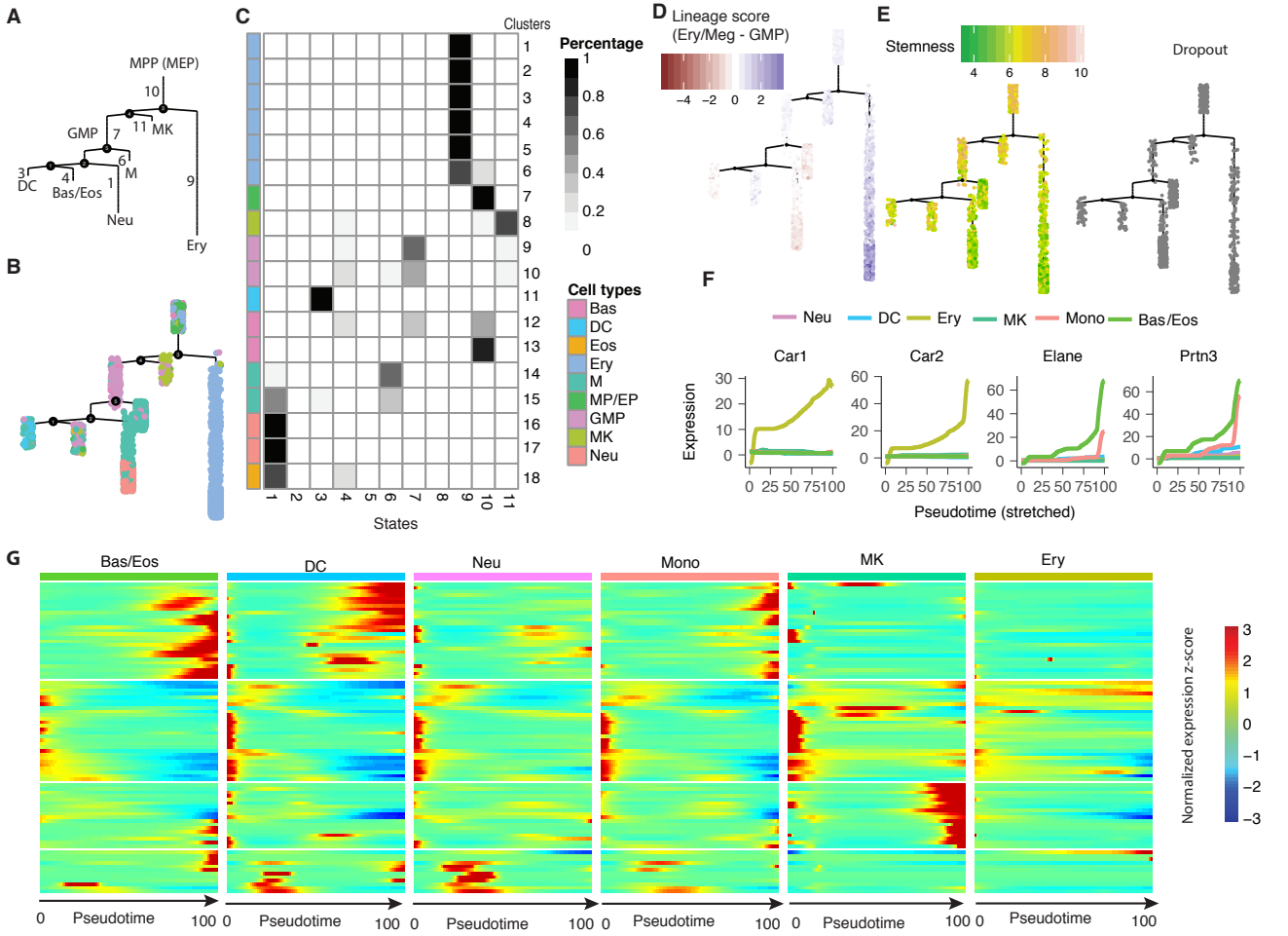


Supplementary Figure 14. Robustness of Monocle 2 under a large range of parameters used in DDRTree. Each panel shows the Pearson correlation of pseudotimes and ARI of branch assignments with respect to the results obtained by Monocle 2 when run as described in the Methods section "*Details on analyzing datasets used in this study*". (**A**) HSMM dataset (**B**) Paul dataset (**C**) Lung dataset (**D**) Olsson dataset. All parameters accepted by DDRTree are included in this analysis.

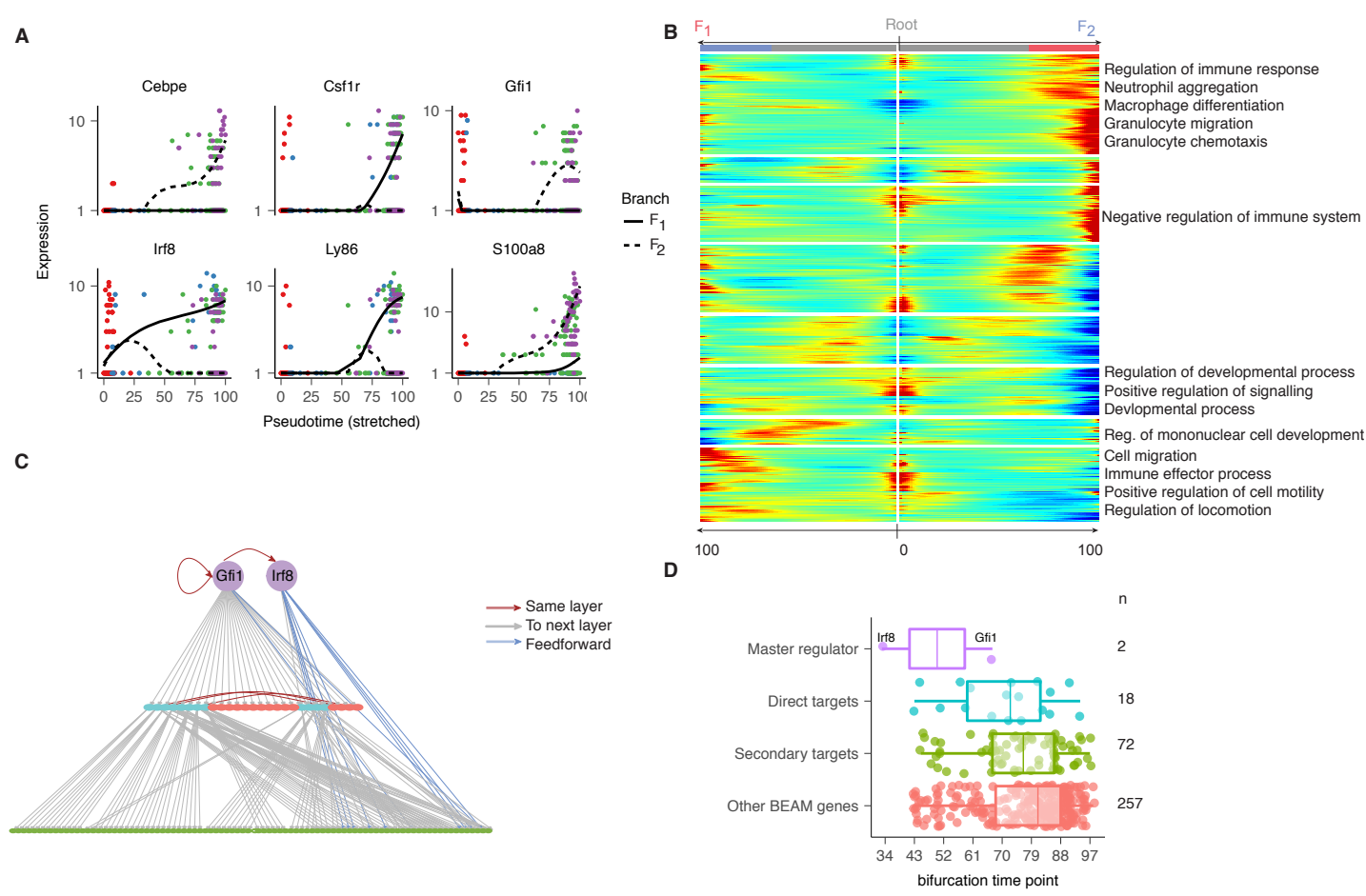




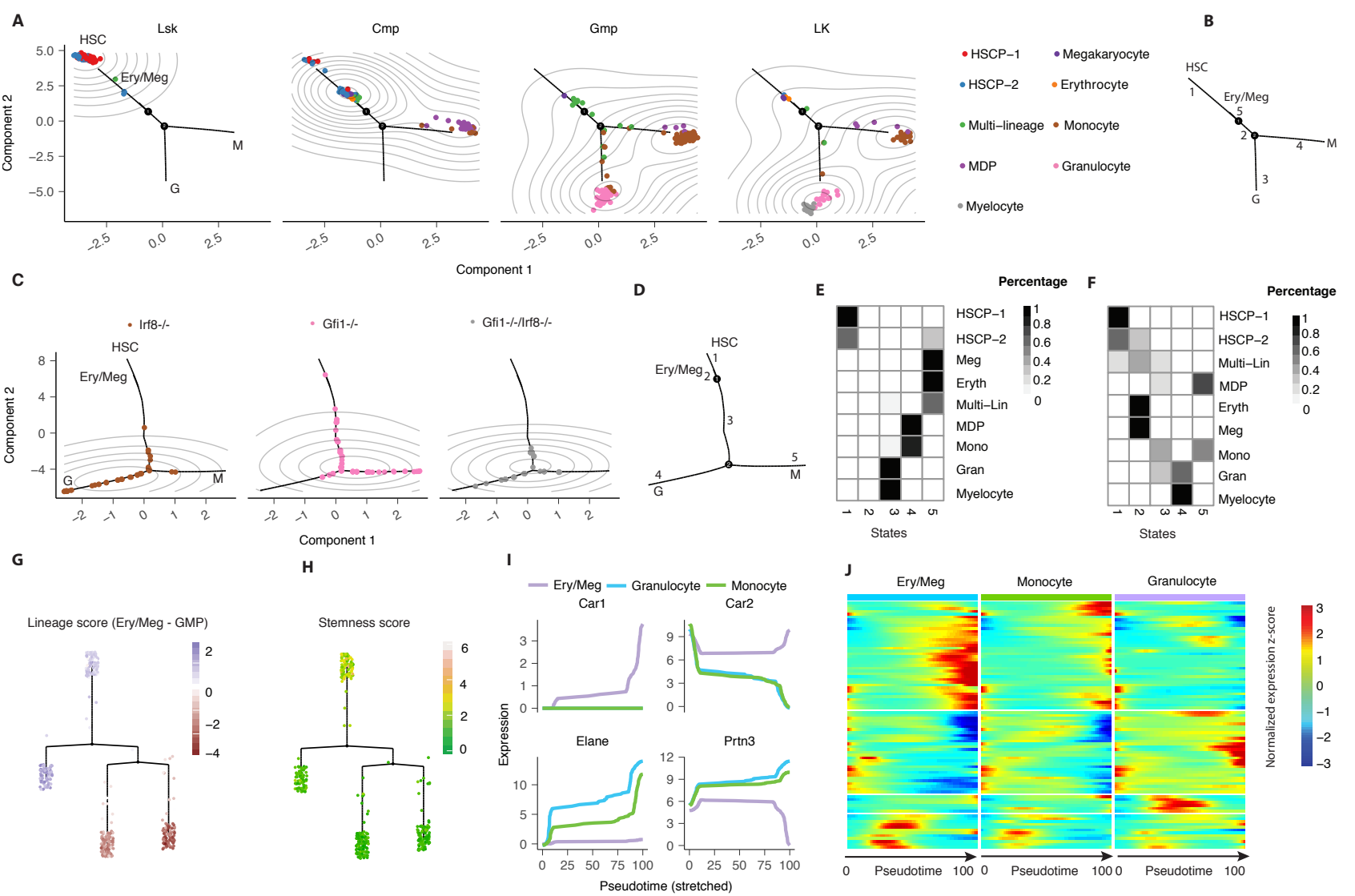
Supplementary Figure 15. Different reversed graph embedding techniques reveal the same incomplete myogenesis branch. (**A**) Trajectory learned with the different RGE techniques, on the (SimplePPT, SGL-tree, L1-graph) or initialized with the (DDRTree) reduced two dimensional space obtained by running PCA, ICA, LLE or diffusion maps. (**B**) *Pearson's Rho* between pseudotime calculated with DDRTree initialized with PCA dimension reduction (reference) and pseudotime calculated with DDRTree, SimplePPT, L1 SGL-tree or L1-graph based on reduced dimension space obtained with different dimension reduction methods, using the dpFeature selected genes as used in **Figure 1B,C** (same as panel **C**). (**C**) Same analysis as in panel B but for branch consistency by Adjusted Rand index. Benchmark for L1-graph is not included in panel **B, C** because it often learns a general graph and awaits for further developments of new pseudotime and branch assignment techniques.



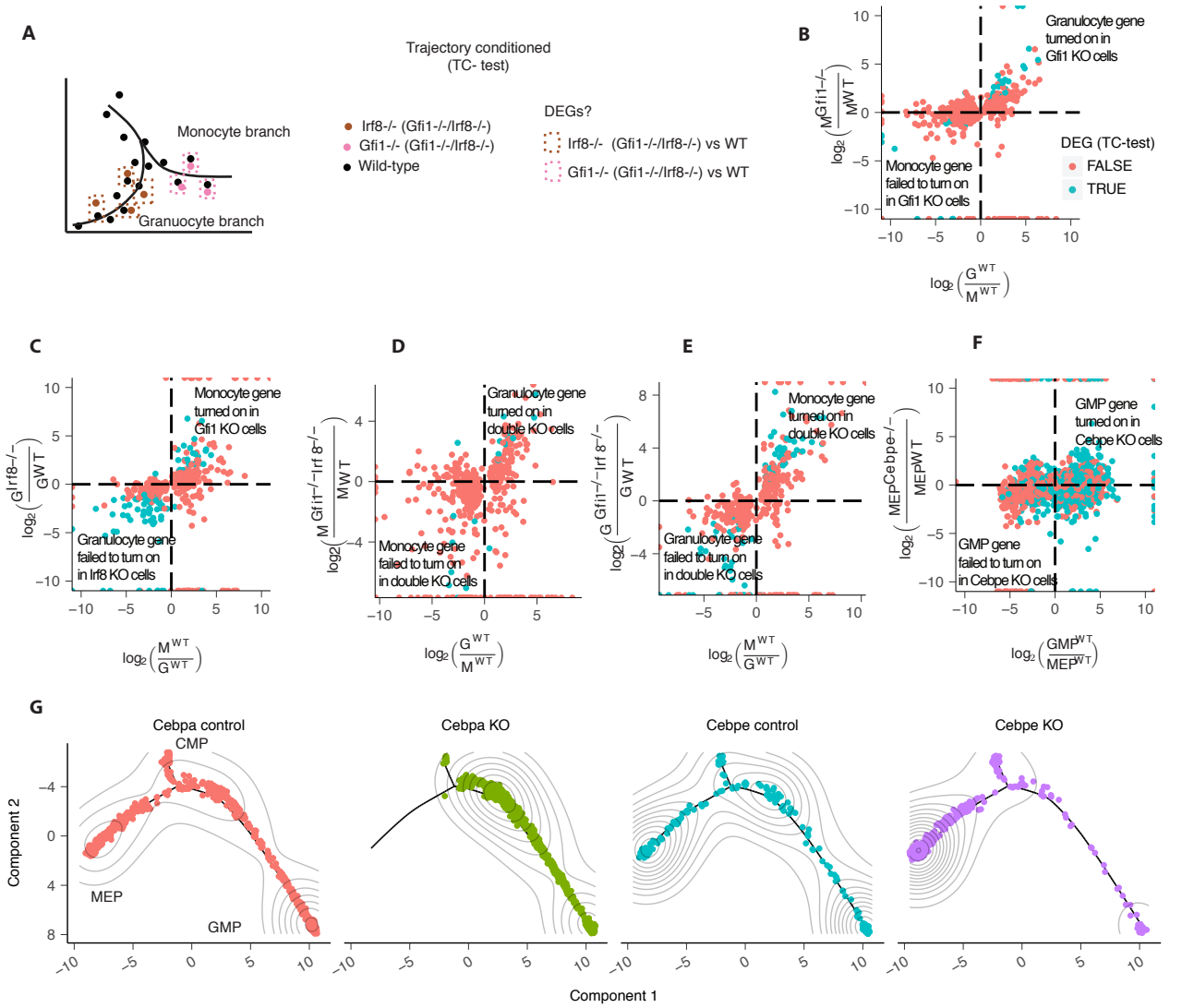
Supplementary Figure 16. Monocle 2 resolves a complex haemopoiesis hierarchy for the Paul dataset. (**A**) The skeleton of the trajectory learned by Monocle 2 describing the lineage relationships learned with Monocle 2. The numerical labels correspond to the "State" label of each segment of the tree. We have added labeled with our interpretation of corresponding cell type (inferred based on comparison with classifications made in the original study, see **B** and **C**). **MPP (MEP)**: multipotent pluripotent progenitor or myeloid and erythroid progenitors; **MK**: megakaryocyte; **GMP**: granulocyte and monocyte progenitor; **DC**: dendritic cell; **Neu/Eos**: neutrophil or eosinophil, **Bas**: basophil, **M**: monocyte; **Ery**: erythrocyte. (**B**) The trajectory learned with Monocle 2 as in panel A, where cells are colored by the cell types suggested by the original study. The trajectory is reconstructed in 10 dimensions but visualized as tree layout in two dimensions using *layout_as_tree()* from the igraph package. (**C**) The distribution of clusters from the original study in each segment of the tree as shown in A, B. c.f. a similar **Figure N4B** of Haghvergdi et al[4]. (**D, E**) Lineage or stemness score for cells on the tree (Same analysis as Olsson dataset, see **Supplementary Figure 18G, H**). Genes used for calculating lineage (**D**) / stemness score (**E**) are based on significant genes returned by differential expression test from the Olsson dataset. Cells show in the "Dropout" panel are those that don't express all genes used in calculating the score. (**F**) Kinetic curves for an example subset of genes used for calculating lineage score in panel **D**. Each curve corresponds to the dynamics of that gene in a particular lineage. (**G**) Multi-way heatmaps for all the genes used in panel **D**. Each sub-heatmap corresponds to a particular lineage where its pseudotime on x-axis starts from 0 (i.e. the root cell). Similar to panel **F**, six lineages are shown.



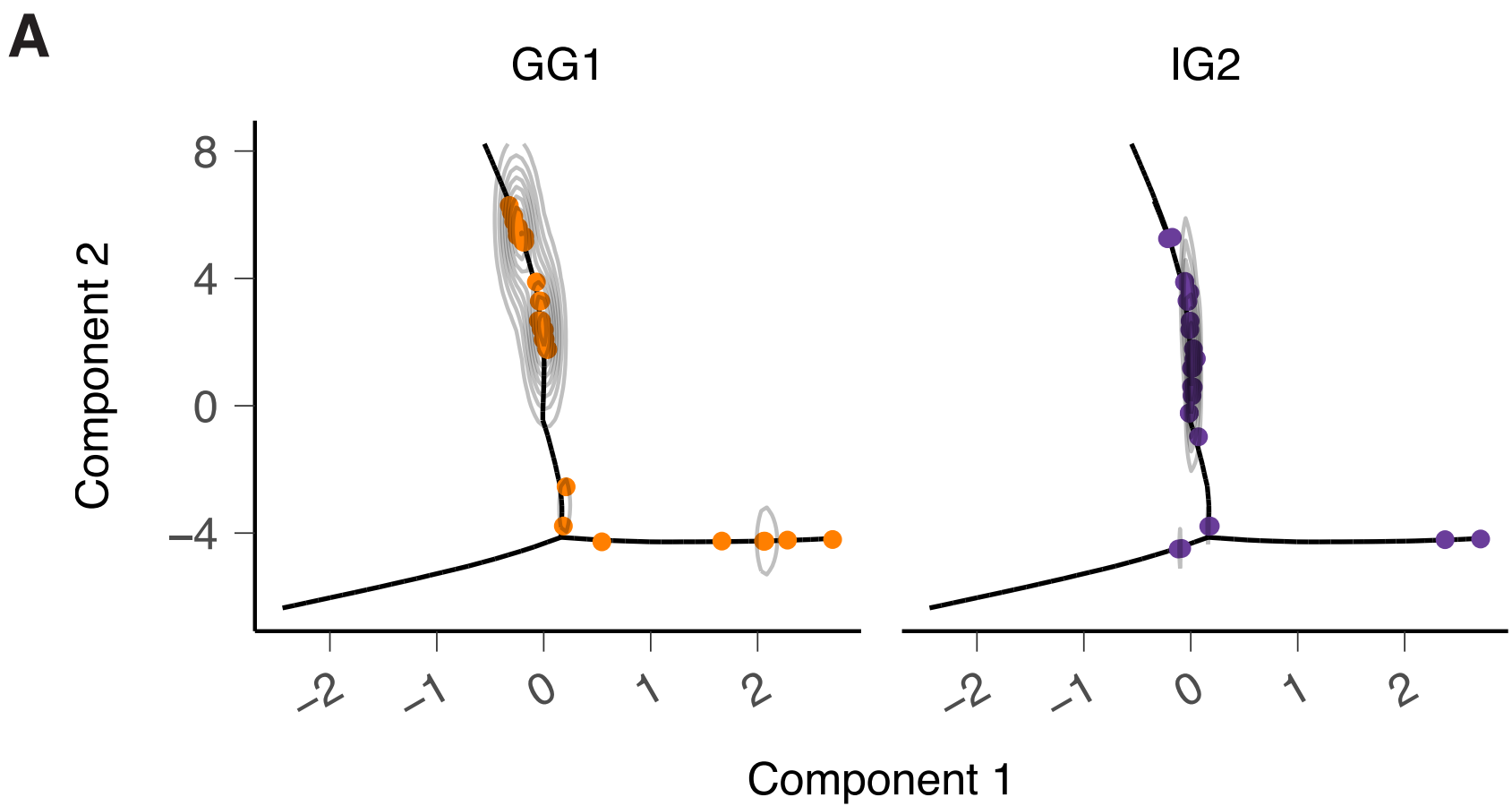
Supplementary Figure 17. Monocle 2 trajectory branches correspond to developmental fate decisions. (**A**) Branch kinetic curves of markers of the monocyte and granulocyte fates. (**B**) Branched heatmap for all the significant branch genes for the wild type data (1015 genes, BEAM test, FDR < 10%). (**C**) A network plot describing direct targets of Irf8 and Gfi1 (derived from ChIP-Seq) and secondary targets (derived by motif analysis; see **Methods**). (**D**) Distribution of bifurcation time points for Irf8, Gfi1, and their direct and secondary targets as well as other genes in panel **C**.



Supplementary Figure 18. Monocle 2 resolves a complex haemopoiesis hierarchy for the Olsson dataset. (**A-D**) Trajectory for the Olsson WT (**A, B**) or full (**C, D**) dataset visualized with component 1 and 2. (**B, D**) are skeletons of the trajectories for the wildtype (WT) (**A**) or full (**C**) dataset and its corresponding states as well as lineages. (**E, F**) Distribution of cells from each cluster from the original study into different branches on the WT dataset (**E**) or full dataset (**F**). (**G, H**) Lineage (**G**) or stemness score (**H**) for cells on the tree for the WT dataset. Genes used for calculating lineage / stemness score are based on significant genes returned by differential expression test from the WT dataset. (I) Kinetic curves for example genes used for calculating lineage score in panel **G**. Each curve corresponds to the dynamics of that gene in a particular lineage. (**J**) Multi-way heatmaps for all the genes used in panel **G**. Each sub-heatmap corresponds to a particular lineage where its pseudotime on x-axis starts from 0 (or the root cell). Similar to panel **I**, three lineages are shown.



Supplementary Figure 19. Trajectories reveal how genetic perturbations divert cells to alternative fates. Panels (**A-E**) are based on the data from Olsson et al while panels (**G, H**) are based on the data from Paul et al. **(A)** A "trajectory-conditioned" test for identifying genes that are differentially expressed between genotypes that controls for the cells' positions on the trajectory. (**B-C**) Expression changes for branch-dependent genes (shown in **Supplementary Figure 17B**) between wild-type granulocytes and monocytes (horizontal axis) plotted against the effects of *Gfi1* or *Irf8* knockout. The vertical axis in panel **B** shows expression changes in *Gfi1-/-* monocytes relative to wild-type monocytes, while panel **C** shows changes in *Irf8-/-* granulocytes compared to wild-type. Genes with significant trajectory-conditioned differences in expression between knockout and wild-type are highlighted (FDR < 10%). (**D-E**) Expression changes in BEAM genes between double KO "monocytes" (**D**) or "granulocytes" (**E**) and wild-type. (**F**) The trajectory-condition test between *Cebpe-/-* cells and nearby wild-type cells on the MEP branch reveals aberrant expression of GMP-branch specific genes in MEP branch. (**H**) Monocle 2 accurately positions cells from *Cebpa-/-* or *Cebpe-/-* collected by Paul et al. Loss of Cebpa fully blocked the MEP branch while *Cebpe* KO partially blocked the GMP branch.



Supplementary Figure 20. Monocle 2 correctly positions transient wild-type cells. (**A**) Gated rare transient cell from Olsson et al are enriched upstream of the branch point separating monocytes and granulocytes.